統計数理(2024)
第 72 巻 第 2 号 233–244

特集「公的統計―社会の重要な情報基盤―」
［研究ノート］

多項ロジットモデルに基づく統計的マッチングの欠測値補完への応用

高部 勲†

(受付 2024 年 2 月 16 日；改訂 7 月 22 日；採択 7 月 25 日)

要　　旨

統計的マッチングは，異なるデータを組み合わせて有用なデータを構築するための手法である．統計的マッチングにより，追加の調査やデータの収集を行うことなく，有益なデータを作成することが可能となり，近年，様々な分野で利用が進んでいる．本研究では，高部・山下(2021)，Takabe and Yamashita (2020)及び高部・山下 (2018)で提案された，多項ロジットモデルに基づく統計的マッチングの手法について紹介するとともに，その副産物として得られるマッチング確率を欠値補完に活用する方法について検討し，実データを用いて，その試算・分析を行った．

キーワード：統計的マッチング，多項ロジットモデル，ウエイト付き距離関数，欠測値補完．

1. はじめに

1.1 公的統計データの欠測値補完に関する概要

社会学におけるアンケート調査，企業による市場調査，行政機関が実施する公的統計調査など，様々な調査においては，無回答・無記入により，調査対象や調査項目の一部についての情報が得られない場合がある．このような欠落した値を欠測値という（日本統計学会, 2022, 2023)．欠測値の要因である無回答には，白紙や未回収など，構成要素単位で発生する全項目無回答(unit nonresponse)と，いくつかの質問に回答がないなど，変数単位で発生する一部項目無回答(item nonresponse)がある（高橋・渡辺, 2017)．欠測値が発生した場合，そのままでは統計的計算処理が不可能となることから，調査対象への再調査等により，可能な限り正しいと考えられる結果を得ることが望ましい．しかし，そのような対応が困難な場合には，欠測値を何らかの方法により補完(imputation)する場合がある（日本統計学会, 2022, 2023)．

単一の統計表の公表を念頭に置いた欠測値の補完においては，1 つの欠測値に 1 つの値(平均値，モデルによる予測値など)を代入する単一代入法(single imputation)が用いられる場合が多い．そのような場合に，得られたデータを用いて関心のあるパラメータの推定を行った場合，例えば分散の過小推定が起こるなど，パラメータの推定値にバイアスが生じる可能性があることが問題点として挙げられる．こうした問題を回避するために，ベイズ統計の考え方を基に，欠測値を含むデータの分布から独立かつ無作為に抽出された複数のシミュレーション値によって欠測値を置き換える多重代入法(multiple imputation)が用いられる場合がある（高橋・渡辺,

† 立正大学 データサイエンス学部：〒 360–0194 埼玉県熊谷市万吉 1700

2017; 高井 他, 2016)．欠測値を含むデータの分析においては，完全にランダムな欠測(MCAR: missing completely at random)，ランダムな欠測(MAR: missing at random)などの欠測のメカニズムを勘案しつつ，推定対象の特性も考慮しながら推定を行う必要があり，そのための様々な手法が提案されている（高橋, 2022; 高橋・渡辺, 2017; 高井 他, 2016; 星野, 2009; 岩崎, 2002)．

ところで，国勢調査のような公的統計調査のデータに欠測値が生じた場合の補完については，我が国を含む多くの国において，公表される多様な集計表の元となる単一の集計用ミクロデータを作成することを念頭に，整合性のあるデータを当該欠測値に代入して補完を行う単一代入法が用いられており，特に同じ調査データの中から欠測した値に類似していると考えられるレコードを当該データから検索して代入するホットデック法(hot deck imputation)や，行政記録などの外部データから検索して代入するコールドデック法(cold deck imputation)が多く利用されている（北原・寺垣内, 2023; 坂下, 2018; 統計委員会担当室, 2013)．

ホットデック法及びコールドデック法では，欠測値を伴うレコードをレシピエント(recipient)と呼び，代入値を提供するレコードをドナー(donor)と呼ぶ．そして，それらのレコードに共通に含まれる変数から何らかの形で距離関数を定義して，距離が最小となるレコードをドナーとする最近隣法(nearest neighbor imputation)が用いることが多い．また，単一のレコードの値を代入するのではなく，複数のレコードの値を何らかのウエイトに基づいて加重平均した値を代入する fractional imputation と呼ばれる手法も考案されている（De Waal et al., 2011)．ホットデック法及びコールドデック法は，2 つのデータの間で類似するレコードを探索するという観点から，統計的マッチングと本質的には同じ考え方であることが指摘されている（高橋・渡辺, 2017)．

本稿では，こうした考え方に基づき，国内外の動向も勘案しつつ，公的統計データを念頭に置いた，統計的マッチング(statistical matching)の手法の欠測値補完への応用について検討する．その準備として次節では，統計的マッチングの概要について説明する．

1.2 統計的マッチングの概要

近年，情報処理技術の発展やネットワーク環境の向上により，様々なデータが利用可能になっており，これらのデータを何らかの形で結合することができれば，新たに統計調査やデータの収集等を行うことなく，情報量の多い有用なデータを構築することができる．

こうした中で，複数のデータを結合するためのデータリンケージの手法が，様々な分野で注目を集めている（Herzog et al., 2007; Christen, 2012; Harron et al., 2015)．ところで，データリンケージを行う際に，各レコードを識別できる照合キー(共通一連番号，名称，所在地など)が存在する場合には，それらを利用した完全照合(exact matching)が可能となる（村田・伊藤, 2016; 山口, 2014)．しかし，このような照合キーが利用できない場合には，各データの共通変数を基に算出した距離が近いレコード同士を結合する方法が用いられる．これを統計的マッチング(statistical matching)という（美添, 2005)．統計的マッチングのイメージを示したものが以下の図 1 である．

統計的マッチングに関する研究は 1960 年代から行われてきており，初期には，名称・所在地などを基に，異なるレコードを同一と判定する確率と，同一の対象を表すレコードが正しく同一であると判定される確率の比率を基にマッチングの適否を判定する方法（Newcombe et al., 1959; Fellegi and Sunter, 1969)が開発されてきている．その後，共通変数以外の変数を欠測値とみなして重回帰モデルやベイズ統計学の枠組みに基づいて推測を行う方法（D'Orazio et al., 2006; Rassler, 2002; 栗原, 2015)，各レコードがどちらのデータに属するかという確率(傾向スコア(Propensity Score))の値が近いレコード同士をマッチングする方法（Rubin, 1986; Stuart, 2010)などの様々な手法が研究・開発されている．

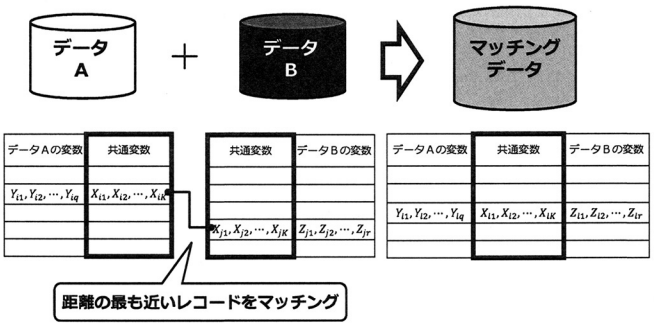

図 1. 統計的マッチングのイメージ.

距離に基づく統計的マッチング(各データに共通の変数を用いてレコード間の距離を計算し，最も近いレコード同士のマッチングを行う手法，D'Orazio et al., 2006)は，比較的初期の段階から研究が行われてきている．ただし，この方法では，各変数の重要度やスケール調整の方法をどのように決定するかについて一般的な基準が無く，各変数のウエイトの決定方法が恣意的になるおそれがある．この問題に対応するため，高部・山下 (2021)，Takabe and Yamashita (2020)及び高部・山下 (2018)では，多項ロジットモデルを用いた統計的マッチングの手法を提案している．この手法は，前述の先行研究における課題を克服している点が特長である．具体的には，距離の最適なウエイトをデータから統計的に(最尤法により)推定することが可能であり，名称・所在地などの詳細な文字情報が利用できない場合でも適用可能な方法であり，さらに，連続変数とカテゴリ変数が混在する場合でも適用可能な方法となっている．また，通常の多項ロジットモデルの推定の際に得られる t 値・p 値などの統計量により，ウエイトの推定精度を分析することが可能であり，レコードの一致確率(マッチング確率)を推定し，マッチングの精度を定量的に評価することが可能である．

1.3 本稿で提案する欠測値補完への応用

本稿では，高部・山下 (2021)，Takabe and Yamashita (2020)及び高部・山下 (2018)における多項ロジットモデルを用いた統計的マッチングの手法について，その概要を説明するとともに，モデルの推定の過程で副産物として得られるマッチング確率の欠測値補完への効果的な利活用方法について提案・検討する．具体的には，データから推定した統計的マッチングに関する多項ロジットモデルを用いて，外部のデータからマッチング確率が最大となるレコードを検索し，そこから値を代入するコールドデック法について検討する．また，本来はホットデック法を念頭に提案されている fractional imputation の考え方を準用し，マッチング確率をウエイトとして用いて，単一のレコードではなく複数のレコードの加重平均により欠測値を補完するコールドデック法の改善方法についても検討する．そして，提案手法を実際のデータ(公的統計データ及び商用データ)を基に，人工的に欠測させた値に対して提案手法を適用し，その効果について検証する．

次節では，提案手法のベースとなる，多項ロジットモデルを用いた統計的マッチングの手法について，高部・山下 (2021)，Takabe and Yamashita (2020)及び高部・山下 (2018)を基に，その手法や推定方法などの概要を説明する．

2. 多項ロジットモデルに基づく統計的マッチング

2.1 手法の概要

以下の 2 種類のデータ(データ A 及びデータ B)の統計的マッチングを行う場合を想定する.

- データ A(マッチング元,ドナー(Donner)):レコード数 M
- データ B(マッチング先,レシピエント(Recipient)):レコード数 N

ここで,データ A の i 番目のレコードと,データ B の j 番目のレコードが同一の対象である確率 P_{ij} を考える(以下,これをマッチング確率という).ここで P_{ij} は,レコード間の距離 D_{ij} を用いて次のように表現できるものとする.

$$P_{ij} = \frac{\exp(-D_{ij})}{\sum_{j=1}^{N} \exp(-D_{ij})} \tag{2.1}$$

距離 D_{ij} の形式については様々なものが考えられるが,以下の絶対値距離(Manhattan 距離)が用いられることが多い.

$$D_{ij} = \sum_{k=1}^{p} \beta_k |X_{ik} - X_{jk}| \tag{2.2}$$

カテゴリ変数(離散変数)に対しては,以下の距離が用いられる.

$$D_{ij} = \sum_{k=1}^{p} \beta_k I(X_{ik} = X_{jk}) \tag{2.3}$$

ここで $I(X_{ik} = X_{jk})$ は,以下のように定義される関数である.

$$I(X_{ik} = X_{jk}) = \begin{cases} 1 & (X_{ik} = X_{jk}) \\ 0 & (X_{ik} \neq X_{jk}) \end{cases} \tag{2.4}$$

距離のウエイト β_k については,連続変数の場合には,レンジ(最大値から最小値を減じたもの)の逆数やデータの標準偏差の逆数,質的変数の場合は 1 が用いられることが多い(Gower 距離,Gower, 1971).

全てのレコードの組合せに対して計算した距離 D_{ij} の値を基に,式(2.1)を用いてマッチング確率 P_{ij} を推定し,その値が最も大きいレコードと結合することにより,統計的マッチングを行うことができる.多項ロジットモデルに基づく統計的マッチングのイメージを示したものが,以下の図 2 である.

2.2 距離のウエイトの推定方法

次に,距離のウエイト β_k をデータから推定する方法について述べる.式(2.1)を基に,対数尤度関数 L を,以下の式(2.5)のように構成することができる.

$$l(\boldsymbol{\beta}) = \log\left(\prod_{i=1}^{M}\prod_{j=1}^{N} P_{ij}(\boldsymbol{\beta})^{\delta_{ij}}\right) = \sum_{i=1}^{M}\sum_{j=1}^{N} \delta_{ij} \log(P_{ij}(\boldsymbol{\beta})) \tag{2.5}$$

$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ は,距離に含まれるウエイトを表している.δ_{ij} は,データ A(Donner)のレコード i と,データ B(Recipient)のレコード j が同一の場合に 1,それ以外の場合に 0 となる変数である.対数尤度関数 L を $\boldsymbol{\beta}$ に関して最大化することにより,ウエイトの最尤推定値 $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$ が得られる.

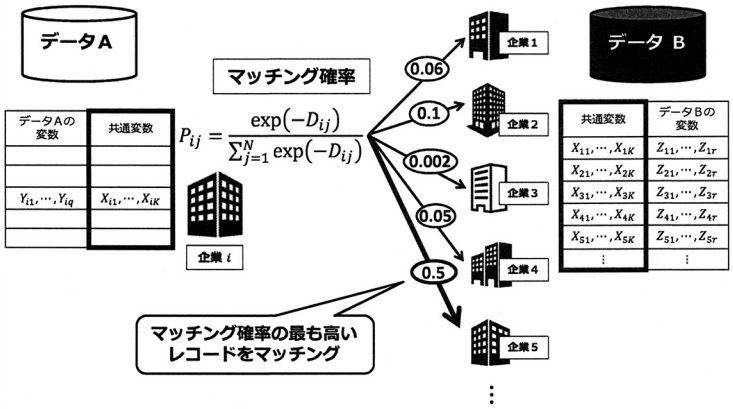

図 2. 多項ロジットモデルに基づく統計的マッチングのイメージ.

前節では，多項ロジットモデルの枠組みに基づき，最尤法により最適な距離のウエイトを推定する方法について述べた．ここでは，距離を用いた統計的マッチングにおいて，距離のウエイトを何らかの形で推定している先行研究とその課題，及びそれらと比較した場合の提案手法の利点について述べる．

2.3　実データを用いた多項ロジットモデルの推定結果

本稿では，平成 24 年経済センサス-活動調査のミクロデータ（経済センサスミクロデータ）及び帝国データバンクの企業データ（帝国データバンクデータ）を対象として分析を行う．推定の対象地域については，3 つの県のデータ（県 A～C）を対象に推定を行う．データの概要については以下のとおり．

経済センサスミクロデータ

- 分析には，平成 24 年調査の結果を使用（調査の期日は平成 24 年 2 月 1 日現在）.
- 一部の変数に関して欠測値が含まれていることから，MICE（Buuren, 2012）の手法に基づき，事前に欠測値を補完（R のパッケージ mice を使用）.

帝国データバンクデータ

- 「COSMOSII」企業概要ファイル・レイアウト C を使用（平成 24 年 2 月時点）.
- 資本金 300 万円以上 5,000 万円未満の企業を対象.
- 完全照合できなかったレコードについては，分析対象から除外.

なお，経済センサスミクロデータと帝国データバンクデータでは，変数の定義などに違いがあり，以下に示すように，各種の調整を行っている．

従業者数及び従業員数

- 帝国データバンクデータの従業員数には，パート・アルバイトを含む場合と含まない場合が混在していると想定されるデータが見受けられ，経済センサスミクロデータの従業者数については，どちらの情報も得られる．
- そこで，上記の 2 つの場合に関して距離を計算し，このうち小さい方を従業者数・従業員数に関する距離とする．

産業
- 帝国データバンクデータの産業分類を，日本標準産業分類を基に組み替えて使用．
- 「S：公務(他に分類されないものを除く)」及び「T：分類不能の産業」は対象外．

開設年
- 帝国データバンクデータでは，開設年が年単位で記録されているが，平成 24 年経済センサス-活動調査では，開設時期がカテゴリ変数となっている．
- そこで，開設年を(1) 1984 年以前，(2) 1985 年〜1994 年，(3) 1995 年〜2004 年，(4) 2005 年以降，の 4 つの時期に区分して，カテゴリ変数として使用．

上記のデータを名称，所在地により完全照合し，各データから 2/3 のレコードを無作為抽出して学習用データとし，残りの 1/3 のレコードを，後述する欠測値補完の検証用のテストデータとした．地域ごとの各データのレコード数については，以下の表 1 のとおり．

経済センサスミクロデータおよび帝国データバンクデータの両方のデータに共通に含まれる変数(共通変数)は，以下の表 2 に示した 6 種類である．

このほかに，売上高も変数として含まれているが，これは後述するように，欠測値補完の検証をする際の変数として用いることとしている．

データには連続変数とカテゴリ変数が含まれていることから，距離関数としては，式(2.2)及び式(2.3)を組み合わせた形のものを用いることとする．その際に，高部・山下 (2021)，Takabe and Yamashita (2020)及び高部・山下 (2018)の結果に基づき，モデルの推定精度を向上させるため，距離の対数変換値(以下の式(2.6))を用いた多項ロジットモデルについて推定を行う．

$$D_{ij} = \sum_{k=1}^{K} \beta_k \log(|X_{ik} - X_{jk}| + 1) \tag{2.6}$$

これらの距離に基づく多項ロジットモデルの推定結果を示したものが，表 3 である．推定結果を見ると，全ての地域で，ほぼ全ての変数について 0.1 パーセントの有意水準で有意となっている．

なお，推定した多項ロジットモデルに基づく統計的マッチングの精度についても重要な観点

表 1．各地域のデータのサイズ．

	地域 A	地域 B	地域 C
(1) 学習用データ	14,744	13,289	18,122
経済センサスミクロデータ	5,095	4,184	5,539
帝国データバンクデータ	9,649	9,105	12,583
(2) テストデータ	7,288	6,646	9,066
経済センサスミクロデータ	2,463	2094	2,774
帝国データバンクデータ	4,825	4,552	6,292
合計 ((1) + (2))	22,032	19,935	27,188

表 2．分析に使用する変数．

変数	単位・区分	種類
従業者数（従業員数）	連続変数	人
資本金額	連続変数	万円
産業分類（大分類）	カテゴリ変数	18 区分
開設年	カテゴリ変数	4 区分
地域	カテゴリ変数	市又は郡に応じた区分
経営組織	カテゴリ変数	株式会社，有限会社又は経営組織不明

表 3. 多項ロジットモデルの推定結果.

変数	地域 A		地域 B		地域 C	
従業員数	1.277	***	1.306	***	1.222	***
	(57.953)		(48.380)		(58.394)	
資本金額	0.816	***	0.859	***	0.775	***
	(76.089)		(58.764)		(69.941)	
産業	3.541	***	3.571	***	3.547	***
	(76.553)		(66.119)		(80.612)	
開設年	1.352	***	1.427	***	1.478	***
	(39.844)		(36.382)		(44.017)	
地域（市・郡）	9.709	***	9.148	***	8.217	***
	(16.787)		(22.354)		(36.774)	
経営組織（株式会社・有限会社）	3.845	***	4.493	***	4.019	***
	(29.731)		(21.094)		(27.863)	
対数尤度	-17221		-11517		-20559	
McFadden の疑似決定係数	0.632		0.698		0.606	

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.01$
( ) は各回帰係数の t 値である.

ではあるものの，今回は，当該モデルを欠測値補完に応用した場合の効果に焦点を当てていることから，マッチングの精度の検証については省略する．様々な距離関数を用いた多項ロジットモデルに基づく統計的マッチングの精度の比較検証に関しては，高部・山下 (2021)，Takabe and Yamashita (2020) 及び高部・山下 (2018) を参照.

3. マッチング確率の欠測値補完への応用

本研究では，多項ロジットモデルに基づく統計的マッチングの手法を欠測値補完に応用する方法について検討する．具体的には，帝国データバンクデータの売上高が全て欠測している場合を想定し，その値を，経済センサス-活動調査データの類似したレコードにおける売上高の値で補完する場合を想定する．なお，本研究では，欠測値があるデータ (レシピエント) を帝国データバンクデータとし，欠測値補完に使用されるデータ (ドナー) を経済センサス-活動調査データとしている[1)].

表 3 における多項ロジットモデルの推定結果を用いて，地域ごとのテストデータにおける欠測値補完の精度を検証する．具体的には，地域ごとにマッチング確率を算出し，レシピエントにおいて最もマッチング確率が大きいレコードの売上高の値を，対応するドナーの売上高の代入値とする方法を検討する．その際に，マッチング確率を用いて，経済センサスの全てのレコードにおける売上高の値を加重平均した結果による欠測値補完ついても検討する．具体的には，ドナー側のデータにおける i 番目のレコードの売上高 (欠測値) を y_i $(1 \leq i \leq M)$ とし，レシピエント側のレコードの売上高の値 z_j $(1 \leq j \leq N)$ とした場合に，マッチング確率 P_{ij} による加重平均により，$y_i = \sum_{j=1}^{N} P_{ij} z_j$ として欠測値補完を行う方法を検討する.

また，これらの提案手法に対する比較手法として，従来の統計的マッチングの研究においてよく用いられている Gower 距離 (式 (2.2) において β_k を X_{ik}, X_{jk} のレンジ (最大値-最小値) の逆数とし，式 (2.3) において β_k を 1 としたもの，Gower, 1971) に基づく単純な最近隣法 (レシピエントにおいて最も Gower 距離が小さいレコードの売上高の値を代入値とする方法) による結果についても検討する.

欠測値補完の精度については，帝国データバンクデータの売上高の値を真値として，以下の平均絶対誤差率 (MAPE: Mean Absolute Percentage Error) を用いて評価する.

$$MAPE = \frac{100}{n} \sum_{i=1}^{M} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \tag{3.1}$$

表 4. マッチング手法ごとの欠測値補完の精度(MAPE).

マッチング手法	地域 A	地域 B	地域 C
Gower 距離に基づく最近隣法	129.06	183.09	111.85
マッチング確率に基づく最近隣法	114.74	147.08	105.57
マッチング確率に基づく加重平均	107.43	133.44	104.78

ここで，y_i は真の売上高，$\hat{y}_i$ は，ドナーとして選択されたレコードの売上高を示している．各地域における欠測値補完の手法ごとに，平均絶対誤差率の結果を示したものが表 4 である．

上記の結果を見ると，いずれの地域においても，Gower 距離に基づく最近隣法よりも，マッチング確率に基づく手法の方が，平均絶対誤差率でみた場合の精度が向上している．また，マッチング確率による加重平均値を用いた手法が，最も精度がよい結果となっている．データから推定された最適な距離に基づくマッチング確率を用いることにより，欠測値補完の精度が向上することが示された．Gower 距離のようにウエイトをデータに応じて調整していない距離を用いるよりも，マッチング確率のようにデータから最適な距離のウエイトを算出している指標を用いる方が，欠測値補完の精度は向上すると考えられる．また，マッチング確率に基づく売上高の加重平均を用いることにより，はずれ値の影響が緩和されると考えられることから，マッチング確率の加重平均値を用いた方が，欠測値補完の精度は向上すると考えられる．こうした結果が，表 4 から確認できる．

なお，地域ごとに平均絶対誤差率の水準に違いがあることが確認される．これについて，平均をとる前の事業所ごとの絶対誤差率を地域ごとにプロットした結果が，図 3，図 4 及び図 5 である．これらの図をみると，特に地域 B において極端に大きな絶対誤差率を持つ事業所が多いことがわかる．このように，2 つのデータの時点などの違いにより，売上高の結果が大きく異なるにもかかわらず従業差数や資本金などの結果が近い値となっている事業所が多く存在する地域においては，他の地域よりも平均絶対誤差率の水準が全体的に大きくなることがわかる．このような場合，地域間の平均絶対誤差率の比較はできないものの，地域内における平均絶対誤差率を基にした各手法の比較・評価は可能である．

4. おわりに

本研究では，高部・山下 (2021)，Takabe and Yamashita (2020) 及び高部・山下 (2018) で提案された多項ロジットモデルに基づく統計的マッチングの手法を基に，マッチング確率を用いた加重平均により欠測値補完を行う方法について提案し，実データを基にその検証を行った．検証の結果，いずれの地域においても，Gower 距離に基づく最近隣法よりも，マッチング確率に基づく最近隣法の方が，平均絶対誤差率でみた場合の欠測値補完の精度が向上しており，データから最適なウエイトを計算した距離関数に基づくマッチング確率は，欠測値補完を行う上でも有用であることが示唆される結果となった．データの構造が大きく変わらないことを想定すれば，一度，距離関数(多項ロジットモデル)を推定しておけば，その結果を用いることで，外部データからのコールドデック法に基づく欠測値補完を効率的に行うことが可能となる．ただし，マッチング確率による加重平均値を用いた手法については，一部の地域で補完の精度が低下することが明らかとなった．売上高は一般的に外れ値を持つことから，加重平均によって補完値の精度が低下した可能性もあり，例えば加重平均を計算する範囲をマッチング確率が上位のレコードに限定するなど，さらに工夫や検証を重ねる必要があると考えられる．なお，本研究では，帝国データバンクデータの売上高が全て欠測している場合を想定しているが，本稿で提案した手法は売上高などの特定の属性値が全てのレコードにおいて欠測している場合に限定されるものではなく，売上高などの属性値が一部のレコードにおいて欠測している場合にも

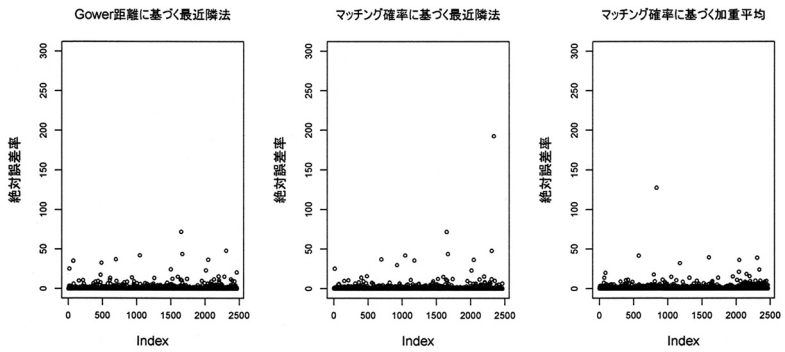

図 3. 地域 A の欠測値補完手法ごとの各事業所の絶対相対誤差.

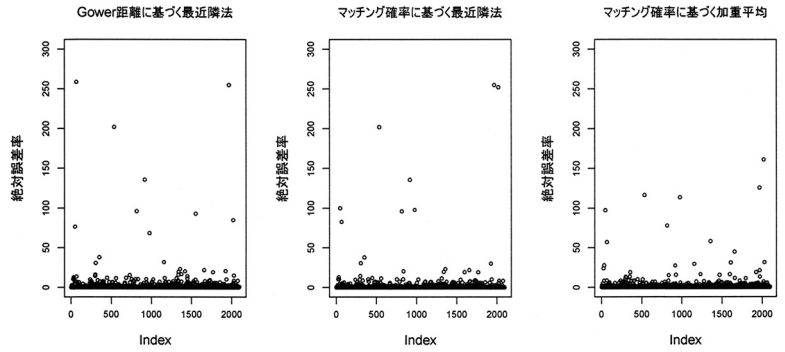

図 4. 地域 B の欠測値補完手法ごとの各事業所の絶対相対誤差.

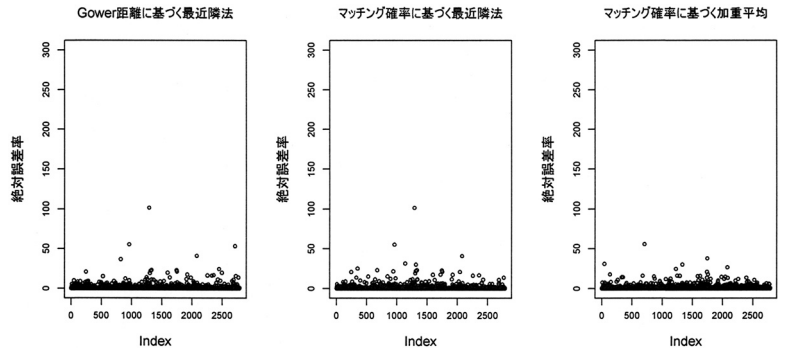

図 5. 地域 C の欠測値補完手法ごとの各事業所の絶対相対誤差.

適用可能である.

今後の課題として，今回のデータを用いて構築したモデルを全く別のデータに適用し，その精度を検証することが考えられる．本研究における手法は，企業データだけでなく，世帯・個人などが対象のデータにも適用することが可能であることから，企業データ以外の様々なデータに対しても本研究における手法を適用することにより，提案手法の有効性を確認していくことが必要であると考える．また，今回は売上高という単一の連続変数について検証を行ったが，複数の変数を対象とする場合の検討や，地域・産業などのカテゴリ変数の欠測値補完をど

のように考えるのかという点も，今後の課題である．

令和 5 年 3 月に決定された政府の第 Ⅳ 期公的統計基本計画において，公的統計ミクロデータの利活用に関しては，調査票情報の利活用に係る安全及び国民の安心確保を図りつつ，学術研究分野からの要請への対応を図っていくため，調査票情報のオンサイト利用を始めとする更なる利便性向上を図ることとされている．こうした状況を鑑み，公的統計のミクロデータや社会調査の個票データなど，レコード単位の様々な統計調査のミクロデータの利活用が進められていく中で，欠測値の補完は今後も一層，重要なテーマになっていくものと考えられ，本稿で提案した手法も含め，今後も，継続的な手法の開発・改善を続けていく必要があると考える．

注．

[1] 本研究では，帝国データバンクデータをレシピエントとし，経済センサス-活動調査データをドナーとしているが，その役割を逆にした(ドナーを帝国データバンクデータとし，レシピエントを経済センサス-活動調査データとした)モデルを想定することも可能である．高部・山下 (2021)では，ドナーとレシピエントの役割を入れ替えたモデルも合わせて活用することにより，統計的マッチングの精度を向上させる方法について分析している．

参 考 文 献

Buuren, S. (2012). *Flexible Imputation of Missing Data*, CRC press, Boca Raton.

Christen, P. (2012). *Data Matching*: *Concepts and Techniques for Record Linkage*, *Entity Resolution*, *and Duplicate Detection*, Springer, Berlin.

D'Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical Matching*: *Theory and Practice*, Wiley, Chichester.

De Waal, T., Pannekoek, J. and Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*, John Wiley & Sons, New Jersey.

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage, *Journal of the American Statistical Association*, **64**, 1183–1210.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties, *Biometrics*, **27**, 623–637.

Harron, K., Goldstein, H. and Dibben, C. (2015). *Methodological Developments in Data Linkage*, Wiley, Chichester.

Herzog, T. N., Scheuren, F. J. and Winkler, W. E. (2007). *Data Quality and Record Linkage Techniques*, Springer, Berlin.

星野崇宏 (2009).『調査観察データの統計科学：因果推論・選択バイアス・データ融合』, 岩波書店, 東京.

岩崎学 (2002).『不完全データの統計解析』, エコノミスト社, 東京.

北原昌嗣, 寺垣内雅子 (2023). 諸外国の国勢調査におけるインピュテーション方法, 統計研究彙報, **80**, 137–162.

栗原由紀子 (2015). 統計的マッチングにおける推定精度とキー変数選択の効果：法人企業統計調査ミクロデータを対象として, 統計学, **108**, 1–15.

村田磨理子, 伊藤伸介 (2016). 事業所・企業系のミクロデータを用いたデータリンケージの可能性：賃金構造基本統計調査を例に, 統計学, **110**, 1–17.

Newcombe, H. B., Kennedy, J. M., Axford, S. J. and James, A. P. (1959). Automatic linkage of vital records, *Science*, **130**, 954–959.

日本統計学会 (2022).『経済統計の実際』, 共立出版, 東京.

日本統計学会 (2023).『調査の実施とデータの分析』, 共立出版, 東京.

Rässler, S. (2002), *Statistical Matching*, Springer, New York.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations, *Journal of Business and Economic Statistics*, **4**, 87–94.
坂下信之 (2018). 諸外国における統計調査の欠測値補完方法の動向と手法の体系について, 総務省統計研究研修所リサーチペーパー, **43**, 総務省統計研究研修所, 東京.
Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward, *Statistical Science*, **25**, 1–21.
高部勲, 山下智志 (2018). 多項ロジットモデルを用いた新たな統計的マッチング手法の提案, 統計学, **115**, 1–16.
高部勲, 山下智志 (2021). 企業データの統計的マッチング及びその精度改善, 統計研究彙報, **78**, 21–40.
Takabe, I. and Yamashita, S. (2020). New statistical matching methods using multinomial logistic regression model, *Advanced Studies in Classification and Data Science* (eds. T. Imaizumi, A. Okada, S. Miyamoto, F. Sakaori, Y. Yamamoto and M. Vichi), 265–274, Springer, Singapore.
高橋将宜 (2022).『統計的因果推論の理論と実装』, 共立出版, 東京.
高橋将宜, 渡辺美智子 (2017).『欠測データ処理』, 共立出版, 東京.
高井啓二, 星野崇宏, 野間久史 (2016).『欠測データの統計科学』, 岩波書店, 東京.
統計委員会担当室 (2013).『統計データの補完推計に関する調査報告書』, 野村総合研究所, 東京.
山口幸三 (2014).『失われし 20 年における世帯変動と就業異動：1991 年〜2010 年のミクロ統計データの静態・動態リンケージにもとづく分析』, 日本統計協会, 東京.
美添泰人 (2005). 統計的照合手法の基礎理論と最近の適用例, 青山経済論集, **56**, 43–71.

Missing Value Imputation Using a Statistical Matching Method Based on a Multinomial Logit Model

Isao Takabe

Faculty of Data Science, Rissho University

Statistical matching is a technique for combining different data to construct useful data. Statistical matching enables the creation of useful data without additional research or data collection and has recently been used in various fields. In this study, we introduce the method of statistical matching based on the multinomial logit model proposed in Takabe and Yamashita (2018, 2020, 2021) and also discuss the use of matching probabilities obtained as a byproduct for missing value imputation.

Key words: Statistical matching, multinomial logit model, weighted distance function, missing value imputation.